Published online 21 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D885–D890***
doi:10.1093/nar/gkn764

NCBI GEO: archive for high-throughput functional genomic data

Tanya Barrett*, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter and Ron Edgar

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892, USA

Received September 26, 2008; Accepted October 6, 2008

ABSTRACT

The Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) is the largest public repository for high-throughput gene expression data. Additionally, GEO hosts other categories of high-throughput functional genomic data, including those that examine genome copy number variations, chromatin structure, methylation status and transcription factor binding. These data are generated by the research community using high-throughput technologies like microarrays and, more recently, next-generation sequencing. The database has a flexible infrastructure that can capture fully annotated raw and processed data, enabling compliance with major community-derived scientific reporting standards such as 'Minimum Information About a Microarray Experiment' (MIAME). In addition to serving as a centralized data storage hub, GEO offers many tools and features that allow users to effectively explore, analyze and download expression data from both gene-centric and experiment-centric perspectives. This article summarizes the GEO repository structure, content and operating procedures, as well as recently introduced data mining features. GEO is freely accessible at http://www.ncbi.nlm.nih.gov/geo/.

INTRODUCTION

The Gene Expression Omnibus (GEO) repository was established in 2000 (1) to host and freely disseminate high-throughput gene expression data. The database is built and maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the National Institutes of Health (NIH) in Bethesda, MD, USA. The data are contributed by the research community, often in compliance with grant or journal directives that require data to be made publicly available, thus allowing others to access, evaluate and re-analyze results.

The three principal objectives of this project are to:

(i) provide a robust, versatile database in which to efficiently store high-throughput functional genomic data;
(ii) offer the simplest submission procedures and formats that support complete and well-annotated data deposits from the research community; and
(iii) provide user-friendly mechanisms that allow users to query, locate, review and download experiments of interest.

The fulfillment of these goals can be assessed in terms of database growth and usage. Today the database holds over 10 000 experiments comprising 300 000 samples, 16 billion individual abundance measurements, for over 500 organisms, submitted by 5000 laboratories from around the world. The database typically receives over 60 000 query hits and 10 000 bulk FTP downloads per day, and has been cited in over 5000 manuscripts.

The 'Omix' division

In recent years, microarray technology has seen an explosion of applications that go far beyond analyzing gene expression levels. Examples of such studies include those that examine genome single nucleotide polymorphism (SNP) and copy number variations (commonly called 'array comparative genomic hybridization', or aCGH), genome–protein binding surveys (commonly called

*To whom correspondence should be addressed. Tel: +1 301 402 4057; Fax: +1 301 480 0109; Email: barrett@ncbi.nlm.nih.gov

'chromatin immunoprecipitation on chip', or ChIP–chip), and various epigenomic factors such as nucleosomal positioning and genome methylation status. Additionally, non-array-based methodologies such as next-generation sequencing are increasingly being applied to such genome-wide investigations (2). Despite the fact that GEO was initially set up to store gene expression data generated by microarrays and serial analysis of gene expression (SAGE), the flexible design of the database allows these non-expression or alternative high-throughput data types to be similarly hosted with little extra effort or overhead. Thus, we have been accommodating to requests to accept such data and extended the minimal standards to fit these types. In fact, at the time of writing, over 15% of the data in GEO are non-expression data. Consequently, the name of the database, Gene Expression Omnibus, has become somewhat misleading and perhaps confusing to users. To address this concern, the non-expression data have recently been placed under a new division called 'Omix', which denotes a *mixture of 'omic* data. Other than a handful of minor issues specific to certain data types, the submission and download procedures and formats for Omix data are largely the same as for GEO. Additionally, equivalent levels of data reporting standards as established for expression data (3,4) are being applied to these other 'omic types. This includes requiring raw data, processed data, protocols and adequate sample and experiment descriptive data. All experiments in GEO and Omix have been newly assigned into broad experiment types, making it much easier for users to locate specific data or technology types.

High-throughput sequence data

GEO recently began processing high-throughput sequence data submissions (5–7). It can be expected that next-generation sequencing technologies will become widely used and have a considerable impact on genome-wide surveys (2). GEO accepts sequence data for studies that examine gene expression, gene regulation, epigenetics, or other studies where measuring sequence abundance is central to the experiment design. GEO hosts the processed and analyzed sequence data, together with sample and experiment metadata; raw data files are brokered to NCBI's Short Read Archive sequence database, ensuring that these sequence data are integrated with NCBI's collection of sequence-specific resources (8).

DATABASE STRUCTURE AND DATA FLOW

As discussed in the Introduction section, the GEO database archives a wide variety of rapidly evolving, large-scale, functional genomic experiment types. These experiments generate data of many different file types, formats and content which consequently present considerable challenges in terms of data handling and querying. The GEO database has built-in flexibility to accommodate diverse data types. Notably, tabular data are not fully granulated in the core database. Rather, they are stored as plain text, tab-delimited tables that have no restrictions on the number of rows or columns allowed. Some columns, however, reserve special meaning, and data from these are extracted to secondary databases and used in downstream query and analysis applications as described in the *GEO DataSets* section. Accompanying supplementary and native file types are linked from each record and stored on an FTP server. Contextual biological and other descriptive metadata are stored in designated fields within database tables with appropriate relations and restrictions.

Submitter-supplied data

The overall structure of the core GEO database remains as described previously (1,9). Briefly, data submitted to GEO are stored in a relational MSSQL database partitioned into three entity types:

A *Platform* record is composed of a summary description of the array and a data table defining the array template. For sequence-based technologies, the Platform lists the sequences detected and identified in that experiment. Each row in the table corresponds to a single feature, and includes sequence annotation and tracking information as provided by the submitter. The table may contain any number of columns allowing thorough annotation of the array. Each Platform record is assigned a unique and stable GEO accession number with prefix GPL. A Platform may reference many Samples that have been submitted by multiple submitters.

A *Sample* record is composed of a description of the biological material and the experimental protocols to which it was subjected, and a data table containing abundance measurements for each feature on the corresponding Platform. The table may contain any number of columns in which to comprehensively present results. The metadata fields may hold very large volumes of text to allow elaborate descriptions of the biological source and protocols. Each Sample record is assigned a unique and stable GEO accession number with prefix GSM. A Sample must reference only one Platform and may be included in multiple Series. A Sample record will typically include supplementary files containing the raw (non-normalized) measured data, linked with the primary record.

A *Series* record defines a set of related Samples considered to be part of a study, and describes the overall study aim and design. Each Series record is assigned a unique and stable GEO accession number with prefix GSE. Series records may contain one or more summary tables and supplementary files.

GEO DataSets

The submitter-supplied objects described above are very heterogeneous with regards to the style, content and level of detail with which the experiments are described. But despite this diversity, all expression-based submissions share a common core set of elements:

(i) sequence identity tracking information of each feature on the Platform;
(ii) normalized expression measurements; and

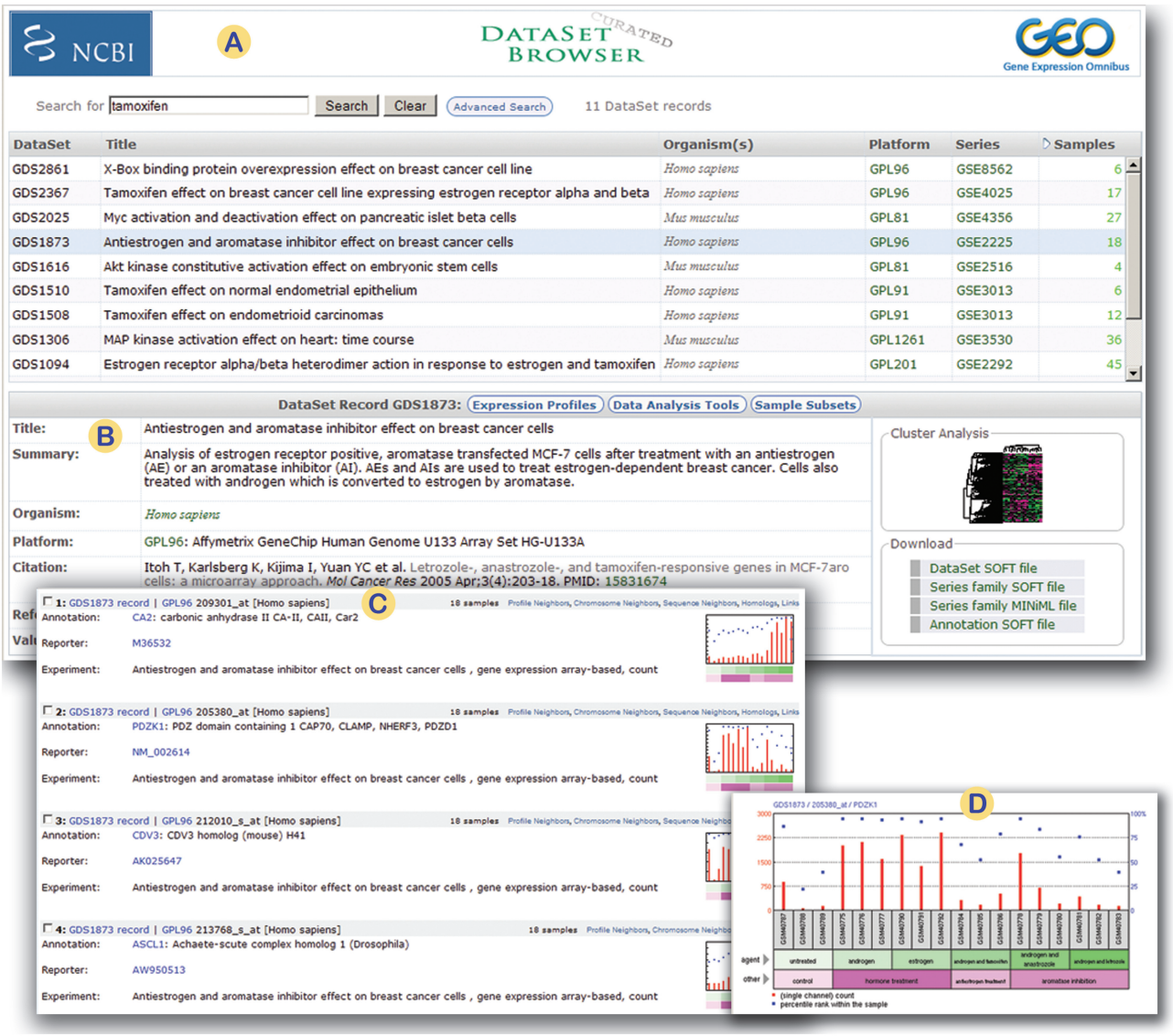

Figure 1. A selection of GEO screenshots. The DataSet Browser (**A**) enables simple keyword searches for DataSets. When a DataSet is selected, a window appears (**B**) which contains detailed information about that DataSet, download options, and links to analysis features including gene expression profiles (**C**). Each expression profile can be viewed in more detail to see the activity of that gene across all Samples in the DataSet (**D**).

(iii) text describing the biological source and experiment aim.

Through a procedure that employs both automated data extraction and manual curation, these three categories of information are captured from the submitter-supplied records and organized into an upper-level object called a GEO DataSet. A DataSet represents a collection of consistently processed experimentally related Sample records, summarized and categorized according to experimental variables. Each DataSet is assigned a unique GEO accession number with prefix GDS. DataSets are a means for transforming diverse styles of incoming data into a relatively standardized format upon which downstream data analysis and data display tools can be built. At this time, only expression-based DataSets are being generated.

DataSets provide two distinct renderings of the data (Figure 1):

(i) an *experiment-centered* representation that encapsulates the entire study. This information is presented as a *DataSet record* which comprises a synopsis of the experiment, a breakdown of the experimental variables, access to auxiliary objects, several data display and analysis tools and download options; and

(ii) a *gene-centered* representation that presents quantitative gene expression measurements for one gene across a DataSet. This information is presented as a *GEO Profile*, and comprises gene identity annotation, DataSet title, links to auxiliary information and a chart depicting the expression level and rank of that gene across each Sample in the DataSet. Gene

annotation is derived from querying sequence identifiers (e.g. GenBank accessions, clone IDs) with the latest Entrez Gene and UniGene databases, an important point given the dynamic nature of gene annotation (10).

SUBMISSION PROCEDURES, FORMATS AND STANDARDS

Great emphasis has been placed on making data deposit procedures as simple as possible for submitters, while not compromising the level of experimental annotation required (11). Four submission options are offered: web forms, spreadsheets, a plain text format and an XML format (Table 1). All these formats are designed to capture all components of the MIAME checklist (4). Deciding which method to use depends on the volume, type of data to be submitted and current data format. To further ease the submission process, native files are requested where possible (e.g. Affymetrix CHP and CEL files). No matter what deposit method is used, the final GEO records will look similar and contain equivalent information. A skilled team of curators is on hand to assist researchers should any questions arise about submission procedures (email: geo@ncbi.nlm.nih.gov).

All data undergo syntactic validation upon upload. A member of the curation team reviews each record to ensure that data are organized correctly and contain sufficient information to interpret the experiment. If content or structural problems are identified, or if critical MIAME components are missing, the curator works with the submitter until the issue is resolved or explained. Submissions are typically approved within 2–5 days, but expedited approval can be performed on request. Researchers are provided the capability to update their records at any time. Records may be kept private until a manuscript describing the data is published. Submitters may generate a temporary reviewer URL that grants anonymous, confidential access to their private data, typically via a journal editor. Guidelines for reviewers and editors regarding how to access and evaluate private data are provided at http://www.ncbi.nlm.nih.gov/projects/geo/info/reviewer.html.

TOOLS TO RETRIEVE, EXPLORE AND VIZUALIZE DATA

Given the wide scope of biological projects and organisms represented in GEO experiments, it is crucial to provide effective query tools so users can quickly locate, analyze and visualize data relevant to their specific interests. A summary of the main query features, and their location and purpose, is provided in Table 2. Figure 2 depicts a schematic overview of the query workflow and how the various features and tools are interlinked.

NCBI's powerful Entrez (PubMed-like) search and linking system serve as the basis for most queries; Entrez GEO DataSets contains experiment-centered data and Entrez GEO Profiles contains gene-centered expression data. Relevant material is located simply by typing in keywords or fielded Boolean (AND, OR and NOT) phrases. Additionally, several auxiliary tools feed into Entrez, including the cluster heat maps and the 'Query group A versus B' tool.

A rich complement of Entrez links is generated to connect data to related information: inter-database links reciprocally connect GEO to other NCBI resources such as PubMed, GenBank and Gene; intra-database links connect genes related by expression pattern, chromosomal position or sequence. Entrez search retrievals can be sorted and filtered by various flags and criteria, and downloaded by various mechanisms. Advanced Entrez features allow generation of multipart fielded queries, or can join multiple queries to identify overlapping results.

GEO provides several graphical renderings that greatly facilitate interpretation and visualization of expression data, including:

(i) interactive pre-calculated hierarchical and on-the-fly *k*-means/*k*-medians cluster heat map images that may hint at groups of coordinately regulated genes.
(ii) Expression profile charts that track the activity of one gene across all Samples in a DataSet. A breakdown of the experimental design is provided along the bottom of the chart, helping the user to quickly assess whether expression levels are shifting with experimental variables.

Table 1. GEO deposit options and formats

Option	Format	Key Points
Web deposit	Web forms	Deposit of individual records. Simple step-by-step interactive web forms.
GEOarchive	Spreadsheets (e.g. Excel)	Batch deposit. Good choice for most users who have many Samples to submit.
SOFT (Simple Omnibus Format in Text)	Plain text	Batch deposit. A simple, line-based, tab-delimited format that can be readily generated, particularly if the data are already in a database.
MINiML (MIAME notation in Markup Language)	XML	Batch deposit. Basically an XML rendering of SOFT format, and similarly suitable if data are already in a database. The XML schema definition is available at the GEO website.

Detailed documentation and example submission templates are available online at http://www.ncbi.nlm.nih.gov/projects/geo/info/submission.html.

(iii) Thumbnail chart images provided on Entrez Profile retrievals that enable rapid visual profile scanning and comparison.
(iv) Value and probability distribution charts that provide rapid indication of how well normalized the data are.

Limited programmatic access is supported using a suite of programs called the Entrez Programming Utilities, or E-Utils. For users who need to perform more robust analyses, all GEO data are available for bulk download via anonymous FTP at ftp://ftp.ncbi.nih.gov/pub/geo/DATA/ and can be imported into external third-party software applications, e.g. the freely available GEOquery package for BioConductor (12).

CONCLUSIONS

NCBI's GEO public archive stores massive volumes of published high-throughput functional genomic data generated by the international research community. In addition to archiving data, tools are provided to assist users of all levels of expertise to quickly search, query, analyze, visualize and download these data. These features employ classic data reduction and filtering methods, succinct displays designed for human scanning, and extensive linking with disparate but related data sources.

Looking at the literature, it is apparent that GEO is used routinely as a primary data resource by the research community. Hundreds of third-party publications cite GEO data as evidence to support or complement independent studies, or use GEO data as the basis of statistical or analytical hypotheses or tools (http://www.ncbi.nlm.nih.gov/projects/geo/info/ucitations.html).

As high-throughput technologies advance, large-scale functional genomic datasets are becoming easier and cheaper to generate. However, a major challenge remains in translating diverse sets of functional genomic data into context, i.e. integrating these datasets with each other and, ultimately, making correlations with observable phenotypes. Collecting and archiving comprehensive 'omic datasets in common formats at one public location like GEO is

Table 2. A summary of the location and purpose of various GEO data mining tools and features

Feature	Purpose
Entrez GEO DataSets http://www.ncbi.nlm.nih.gov/sites/gds	Query interface that facilitates identification of experiments of interest using keywords or fielded Boolean phrases.
Entrez GEO Profiles http://www.ncbi.nlm.nih.gov/sites/geo	Query interface that facilitates identification of gene expression profiles of interest using keywords, gene names, gene symbols, etc. or fielded Boolean phrases.
DataSet browser (NEW) http://www.ncbi.nlm.nih.gov/sites/GDSbrowser	A simple tool to browse, query and filter DataSets.
Advanced Entrez features toolbar at head of all NCBI Entrez query and retrieval pages	'Preview/Index' lists fields and terms by which data are described, assisting users to construct complex multipart queries. 'History' recalls results of previous searches and allows users to combine previous queries to form a new query. 'Display' retrieves related data in other NCBI Entrez resources in batch mode.
Profile neighbors link at top right side of Entrez GEO Profiles retrievals	Connects groups of genes that have similar expression profiles within a DataSet.
Sequence neighbors link at top right side of Entrez GEO Profiles retrievals	Connects groups of genes related by nucleotide sequence similarity across all DataSets.
Homolog neighbors link at top right side of Entrez GEO Profiles retrievals	Connects groups of genes related by HomoloGene groups across all DataSets.
Chromosome neighbors (NEW) link at top right side of Entrez GEO Profiles retrievals	Connects the 20 physically closest genes from each side of the query gene within a DataSet.
Download profile data (NEW) button at the top of Entrez GEO Profiles retrievals	Downloads value data and gene annotation for retrieved profiles.
Links link at top right side of Entrez GEO Profiles and Entrez GEO DataSets retrievals	Connects GEO data to related data in other NCBI Entrez resources, including PubMed, GenBank, Gene, UniGene, OMIM and others.
Subset effect flags intrinsic to Entrez GEO Profiles retrievals and specifiable using [Flag] qualifiers in Entrez GEO Profiles	Identifies genes that display marked differences in expression level between experimental variables. Retrievals are default ordered according to presence of these flags, which increases visibility of potentially interesting genes.
Find gene in this DataSet (NEW) box on Dataset records	A quick way to link to a gene profile of interest within a specific DataSet.
DataSet heatmaps on DataSet records	Interactive images of precomputed hierarchical clusters, user-defined *k*-means clusters, and chromosomal location heatmaps (NEW) that allow visualization, selection and download of clusters or regions of interest.
Query group A versus B on DataSet records	Identifies gene expression profiles that meet user-defined statistical differences (*t*-test or fold difference) between two specified groups of Samples within a DataSet.
Experiment type categories (NEW) on Series and DataSet records and specifiable using [DataSet Type] qualifier in Entrez GEO DataSets	Classification of experiment types into broad categories helps users find relevant data.
GEO blast http://www.ncbi.nlm.nih.gov/geo/query/blast.html	Retrieves sequences and corresponding gene expression profiles that are related by nucleotide sequence similarity to a user-defined sequence.

Features introduced within the last 2 years are labeled NEW.

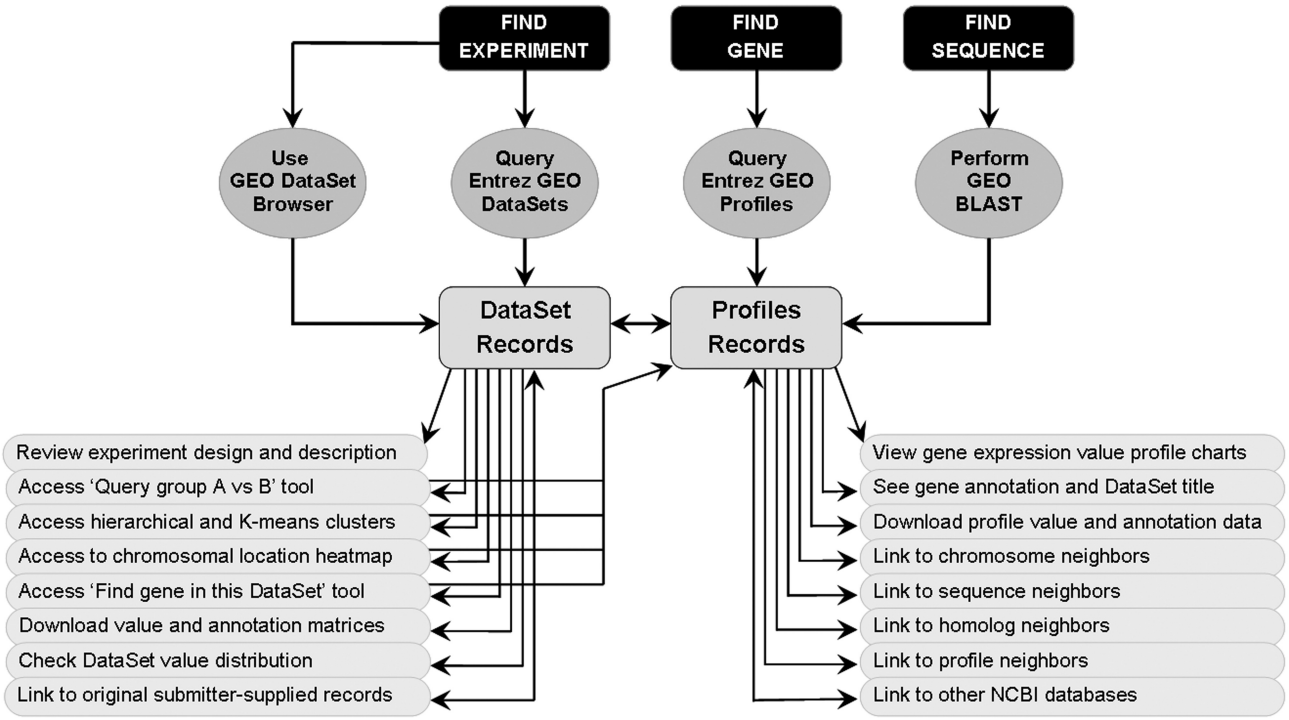

Figure 2. A schematic overview of query workflow, and how various features and tools are interlinked. A description of the location and purpose of many of these features is provided in Table 2.

an important first step in facilitating such large-scale integrative analyses. It can be anticipated that a continued increase in availability of these datasets will contribute to our understanding of how the genome regulates and specifies cellular types, states and processes.

The GEO database and tools continue to undergo intensive development and modification aimed at enhancing the experiences of both data submitters and data consumers. The submission pipeline and data transfer mechanisms will continue to be upgraded, and we plan to develop additional data retrieval and mining features, particularly for the novice user.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health; National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
2. Wold,B. and Myers,R.M. (2008) Sequence census methods for functional genomics. *Nat. Methods*, **5**, 19–21.
3. Ball,C., Brazma,A., Causton,H., Chervitz,S., Edgar,R., Hingamp,P., Matese,J.C., Parkinson,H., Quackenbush,J., Ringwald,M. *et al.* (2004) Standards for microarray data: an open letter. *Environ. Health Perspect.*, **112**, A666–A667.
4. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
5. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
6. Ghildiyal,M., Seitz,H., Horwich,M.D., Li,C., Du,T., Lee,S., Xu,J., Kittler,E.L., Zapp,M.L., Weng,Z. *et al.* (2008) Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, **320**, 1077–1081.
7. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
9. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
10. Barrett,T. and Edgar,R. (2008) Reannotation of array probes at NCBI's GEO database. *Nat. Methods*, **5**, 117.
11. Edgar,R. and Barrett,T. (2006) NCBI GEO standards and services for microarray data. *Nat. Biotechnol.*, **24**, 1471–1472.
12. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.